Exploring the Taxi and Uber Demands in New York City: An Empirical Analysis and Spatial Modeling

Diego Correa Ph.D. [a, *], Kun Xie Ph.D. [b], Kaan Ozbay Ph.D. [a]

[a] *Department of Civil and Urban Engineering, New York University, 6 MetroTech Center, Brooklyn, NY 11201, USA.*
[b] *Department of Civil & Environmental Engineering, Old Dominion University (ODU), 129C Kaufman Hall, Norfolk, VA 23529, USA.*

* *Corresponding author.*
E-mail addresses: dcorreab@nyu.edu (D. Correa), kxie@odu.edu (K. Xie), kaan.ozbay@nyu.edu (K. Ozbay).

ABSTRACT

This study aims to investigate the impact of the emerging app-based for-hire vehicles on taxi industry through quantitative analyses of Uber and taxi demands for neighborhoods of New York City (NYC). Demand forecasting models, which can account for the spatial dependence of Uber and taxi trips are developed. In the empirical analysis, we explore the spatio-temporal patterns of Uber and taxi pick-up data. A high correlation between taxi and Uber pick-ups can be observed, especially in the central areas of the City. From 2014 to 2015, Uber trips increased dramatically by 10 million (223.3%), while taxi trips (include both yellow and green taxis) decreased slightly by 0.8 million (1.0%). The rate of growth of Uber is the lowest in Manhattan (201.2%), and the highest in the outer boroughs like Bronx (597.0 %) and Staten Island (573.0%). Results of the Moran's *I* tests confirm the spatial dependence of both taxi and Uber demands. Linear models, spatial error models, and spatial lag models are developed to estimate the taxi and Uber demands of each neighborhood using socio-economical and transportation-related characteristics. The spatial error models are found to outperform the other two by capturing the spatial dependence via a spatially lagged dependent variable. Neighborhoods with lower transit access time (TAT), higher length of roadways, lower vehicle ownership, higher income and more job opportunities are associated with higher taxi/Uber demands.

Keywords: Uber, taxi, demand, spatial dependence, spatial lag model, New York City

INTRODUCTION

With the rapid growth of app-based car services like Uber and Lyft, the for-hire vehicle industry is changing dramatically. This type of electronic-dispatch on-demand car services provide a more convenient and flexible way of traveling and serve an increasing number of users in recent years *(1)*, especially in big cities. How have the transportation systems been changed and what will be the long-term impact of this change are some of the essential questions for researchers, planners and policy makers to answer. New York City (NYC), the most populous city in U.S., has more than 8.5 million residents *(2)*. NYC taxis carry 172 million trips annually, which account for 11% of all fare-paying riders. Thus, taxi is clearly an imperative transportation mode in the city *(3)*. As of 2015, there are 13,587 licensed yellow taxis and 7,676 green taxis *(4)*, which are managed by the NYC Taxi & Limousine Commission (TLC). The difference between yellow and green taxis is that green taxis are only allowed to pick-up street hails in outer boroughs *(5, 6)*, while yellow taxis can provide service in the whole city. The taxi industry in NYC also faces stiff competition from app-based car services, especially the service provided by Uber *(1)*.

We aim to study of the impact of the emerging app-based for-hire vehicles on NYC taxi industry. There are few studies dealing with this emerging topic, mainly due to the data availability. Recently, as datasets of taxi and Uber trips are made open to the public (the data sources can be found in the data preparation section of this paper), it provides new opportunities to start this research. In this study, we first explore the spatio-temporal patterns of the demand for Uber and taxi. We explore answers to questions such as how the demand for Uber and taxi is changing over time, how is the demand for taxi and Uber distributed over space, what is the relation between the demand for Uber and taxi trips, and what are the factors affecting the demand for Uber and taxi. Understanding these demand related questions may provide useful insights that can guide future planning and regulation efforts for all types of for-hire vehicles in NYC and possibly other cities in the US.

The following section reviews previous studies on taxi demand and modeling. It is followed by the empirical analysis of the spatio-temporal patterns of the Uber and taxi pick-up data. Then statistical spatial models are developed to estimate the demand for Uber and taxi using socio-economic and transportation-related characteristics of the study area. Next, methodology adopted for the estimation of spatial models, spatial dependence tests, modeling results and discussion are presented. This paper ends with conclusions and suggestions for future research.

LITERATURE REVIEW

Taxi complements other public transport modes with a flexible door-to-door service *(7, 8)*. A study conducted in Taipei City showed that 60–73% of their operation hours, taxi drivers were driving without passengers because they did not know where potential customers were, leaving them no other choice than wandering around the city *(9)*. There is another study that applied time series forecasting techniques to real-time vehicle location systems in taxis to make short-term predictions of the passenger demand in the city of Porto, Portugal *(10)*. A predictive model for the number of vacant taxis in a given area based on the time of day, day of the week, and weather conditions in Lisbon, Portugal, is presented *(11)*.

An extensive variety of spatial information sources such as GPS have recently emerged. A GPS based system is also utilized to track all New York City taxis. Various recent research studies used this data source to analyze different aspects of taxi ridership in NYC. One recent study *(12)* analyzed travel times and found that travel times from truck and taxi GPS data can be better

estimated during AM and PM periods than during night time, which indicates that speed differences between taxis and trucks are greater for free-flow conditions. Research on modeling the variation of taxi pick-ups was developed using Poisson *(13)* and negative binomial *(14)* models, have been applied in using NYC taxi data. The model suggests that adjacent census tracts have correlated residuals, meaning that spatial autocorrelation exits.

Other studies that utilized NYC taxi trip data estimated a binary logit model to model the mode choice between transit and taxi modes *(15),* compared trip characteristics between summer (July) and non-summer (March) months *(16),* and developed a data visualization tool namely, TaxiVis, which is a software implementation that allows users to visually query taxi trips by considering spatial, temporal, and other constraints *(17)*.

Another study used ten-month NYC taxi trip data from 2010 to estimate a multiple linear regression model for each hour of the day to model NYC taxi pick-ups and drop-offs *(18)*. The results identified six important explanatory variables for taxi trips, which include population, education, age, income, transit access time, and employment, where the influence of these factors on taxi pick-ups and drop-offs changed at different times of the day. To model spatial variation of taxi trip demand and supply in NYC, Poisson-Gamma-Conditional Autoregressive (CAR) model is developed using a ten-month taxi data set in NYC York City *(19)*. The errors of the CAR model provide insights into when and where there are insufficient taxi supply or surplus taxi supply relative to taxi demand.

In spatial analysis, spatial dependence can be modeled in two ways: using an error term and using a spatially lagged dependent variable *(20).* The former way is referred to as the spatial error specification that assumes the spatial dependence is only due to spatial error correlation effects *(21)*. The latter is denoted as the spatial lag specification which allows spatial dependence through both spatial error correlation effects and spatial spillover effects. An appropriate consideration of spatial dependence can help adjust the effects of casual factors in the statistic models.

Another study was performed to develop an incident duration model *(22)*, which can account for the spatial dependence of duration observations, to investigate the impacts of a hurricane on incident duration. Moran's I statistics *(23)* confirmed that durations of the neighboring incidents were spatially correlated. Moreover, Lagrange Multiplier tests suggested that the spatial dependence should be captured in a spatial lag specification. A spatial error model, a spatial lag model, and a linear model without consideration of spatial effects were estimated for incident duration.

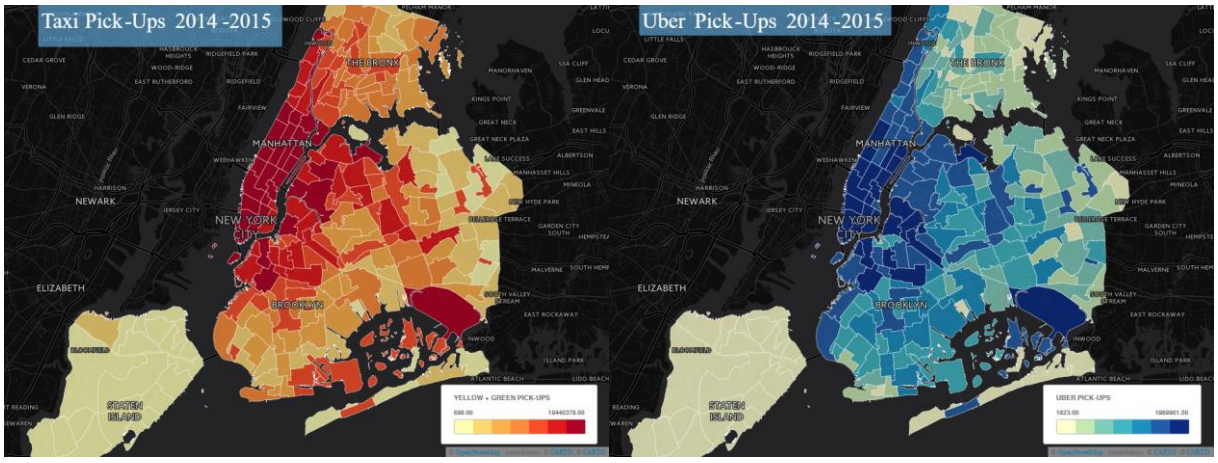
Previous studies developed spatio-temporal models for taxi demand, but detailed Uber pick-up data were not used in those studies, thus the relationship between Taxi and Uber in NYC has not been explicitly investigated. In other words, previous studies didn't investigate the effects of Uber on the overall taxi demand. Moreover, those studies didn't use spatial error and spatial lag models to account for spatial correlation between dependent and independent variables. In this paper, we intent to fill these gaps using Uber and taxi data from NYC.

DATA PREPARATION

We use New York City (NYC) as the study area, including five boroughs: Manhattan, Brooklyn, Queens, Bronx and Staten Island. Neighborhood Tabulation Areas (NTAs) are used as the basic units of analysis. NTAs are composed of census tracts, so it is relatively easy to obtain socio-economic features at the NTA level. There is a total of 195 NTAs in NYC. The geographic information system (GIS) data of NTAs are provided by New York City Department of City Planning NYCDCP *(24)*. The NTAs can be easily connected to the census data provided by U.S. Census Bureau *(25)*. Demo-economic features such as population, employment, and income of NTAs were also obtained. Description and descriptive statistics of key variables are presented *in Table 1*. More details on obtaining taxi and Uber pick-ups as well as transit access time (*TAT*) are presented in the following two subsections.

Taxi and Uber Data

The New York City Taxi & Limousine Commission has released a detailed historical dataset *(26)* covering over 172 million yellow and green taxi trips from April to September in 2014 and from January to June in 2015. Each trip record contains precise coordinates of pick-up and drop-off locations, timestamps for when each trip started and ended, and other variables including fare amount, payment method, and distance traveled. These months was selected to match with the available Uber data during the same period of time. We extracted 19 million Uber rides from Todd W. Schneider's GitHub repository *(27)*. Less detailed than the taxi data, the times and locations are available only for Uber pick-ups. The number of pick-ups and drop-offs was obtained at the NTA level using spatial processing tools, as presented in the figure below. *Figure 1* (a) and (b) present the taxi and Uber pick-ups by NTAs, demands are spatially clustered.



(a) Taxi (b) Uber

Figure 1 New York City Uber and taxi pick-ups visualized by NTA.

Transit Access Time (TAT)

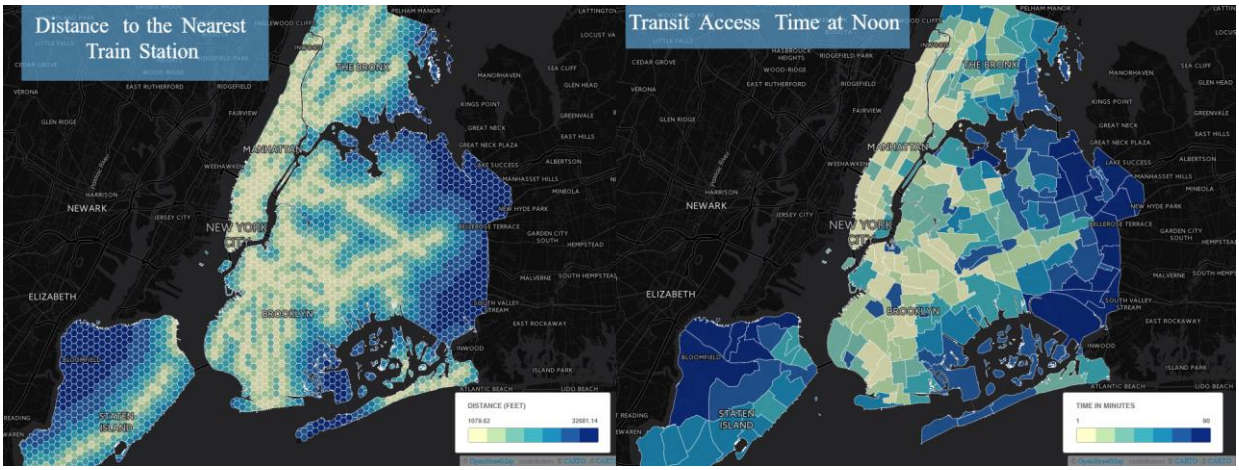
The location of NYC subway stations is available in ESRI's Geodatabase format *(28)*. The geodatabase contains the station geometries for the NYC subway system as well as three commuter rails (Long Island Rail Road, Metro-North Railroad, and Staten Island Railway) within NYC boundaries. Transit data are made available through General Transit Feed Specification (GTFS) files from The Metropolitan Transportation Authority (MTA) *(29)*. GTFS defines a common format for public transportation schedules and associated geographic information. It is used to estimate the frequency of the subway service in time and space. Transit Access Time (*TAT*), the combined estimated walking time a passenger must spend to access the nearest station plus the time that passenger will wait for transit service, was obtained for each NTA. It should be noted that only the subway accessibility is considered in this study. Lower *TAT* value indicates higher transit accessibility. *TAT* is estimated as follows (*18*):

$$TAT = \frac{60D}{v_w} + \frac{60}{f} \quad (1)$$

Where:

- D is the distance for a person at a specific location to the nearest transit stop. (mi)
- v_w is the walking speed set to 3.1 (mi/h) *(30)*.
- f is the frequency of transit services (trains/h)

Hexagonal cells are used to compute *TAT* as shown in *Figure 2* (a). The primary advantage of a hex map over a traditional square grid map is that the distance between the center of each hex cell and the center of all six adjacent hexes is constant. The edge of each cell is 1750 feet, which is small enough since the walking time to cross each cell is less than 6 minutes. Each cell is characterized by the location of its centroid. For each cell the minimum *TAT* was calculated using the distance and waiting time to the near transit stop, that was estimated using k-nearest neighbor's algorithm. Once the minimum *TAT* for each cell is determined, it is not difficult to calculate *TAT* by averaging the values across the cells included within the NTA *Figure 2* (b).



(a) Hexagonal cells (b) NTA

Figure 2 Transit access time (TAT) for each zone in New York City.

All variables presented in this study are described below:

Table 1 Description and Descriptive Statistics (N=195)

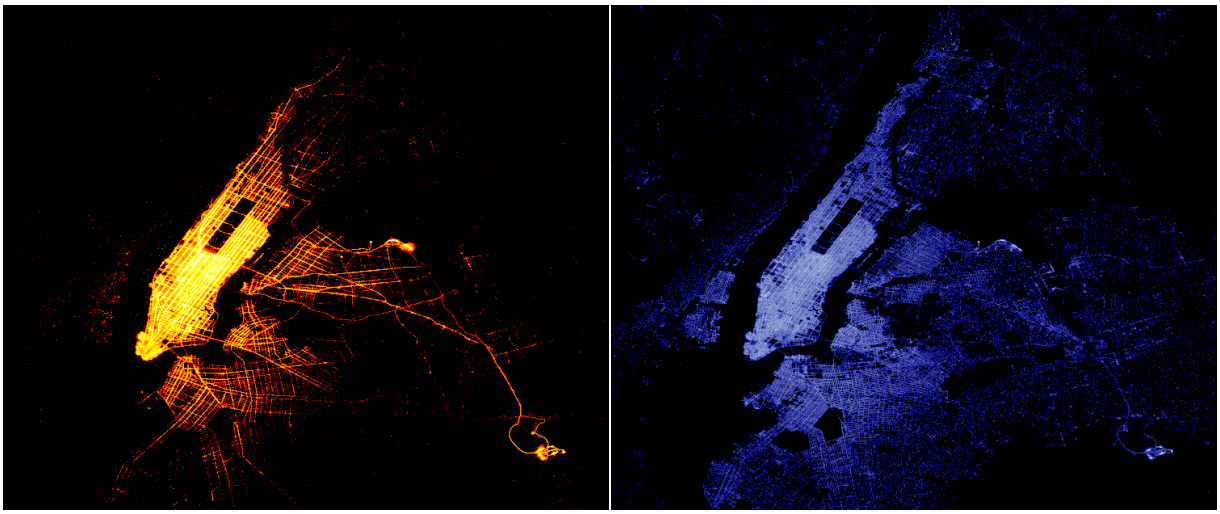
Variable	Description	Mean	S.D.	Min.	Max.
Dependent Variable					
Uber pick-ups	Number of Uber pick-ups (10^3)	68.11	213.08	0.02	1969.90
Taxi pick-ups	Number of Taxi pick-ups (10^3)	627.43	2194.72	0.00	20156.27
Input Variables					
TAT	Transit access time at specific hour (min)	16.77	16.71	1.20	89.90
Pop2010	Total Population in 2010 (10^3)	42.48	22.13	0.42	135.72
EduBac	Population with bachelor's degree or higher (10^3)	9.80	10.09	0.09	77.53
CapInc	Mean household income (10^3 $)	75.54	34.51	28.53	229.92
TotJob	Employed (10^3)	19.67	11.25	0.11	76.27
Age					
PopAge0_14	Population ages below 15 (10^3)	7.69	4.56	0.06	34.47
PopAge15_34	Population ages between 15 and 34 (10^3)	13.30	7.51	0.10	36.39
PopAge35_64	Population ages between 35 and 64 (10^3)	16.38	8.74	0.11	58.45
Housing					
House_1	housing 1 unit (10^3)	2.86	2.65	0.01	12.19
House_2	housing 2 units (10^3)	2.38	2.16	0.01	14.90
House3_4	housing 3 or 4 units (10^3)	1.81	1.96	0.01	8.87
House_5plus	housing 5 or more units (10^3)	10.64	10.87	0.02	71.75
Employment Groups					
JobRet	Retail trade (10^3)	1.96	1.07	0.01	5.47
JobInf	Information (10^3)	0.75	0.85	0.02	5.63
JobFin	Finance and insurance, and real estate (10^3)	2.01	1.96	0.01	12.71
JobPro	Professional, scientific, and management (10^3)	2.45	2.31	0.02	17.13
JobEduc	Educational, health and social assistance (10^3)	5.16	2.99	0.02	19.55
JobFod	Arts, entertainment, recreation, and food services (10^3)	2.01	1.60	0.03	8.57
Vehicles					
NoVeh	No vehicles available (10^3)	8.90	7.73	0.08	50.96
Veh_1	One vehicle available (10^3)	4.99	2.78	0.03	18.03
Veh_2	Two vehicles available (10^3)	1.68	1.31	0.01	6.04
Veh_3	Three or more vehicles available (10^3)	0.49	0.48	0.01	2.65

Transit					
BusStop	Number of Bus Stops	68.07	40.63	1.00	220.00
SubwayStop	Number of Subway Stops	3.71	3.19	1.00	22.00
DistToSubway	Walking Distance to Near Subway Stop (miles)	0.77	0.89	0.06	4.64
FreqTrain	Frequency of trains (trains/hour)	68.50	84.13	5.00	678.00
AreaSQM	Area of individual NTA zone (mile^2)	1.55	1.60	0.20	11.76
Bike					
BikeRack	Number of Bike Racks	94.03	163.00	1.00	870.00
BikeLineLen	Length of bike lines inside NTA (miles)	4.25	5.22	0.02	41.16
Schools	Number of Schools	14.58	9.79	2.00	71.00
RoadsMi	Length of roads inside zone (miles)	48.08	28.36	6.41	203.51
TAT_AMPeak	Transit access time at morning peak period (min)	15.84	16.93	1.10	89.90
TAT_Noon	Transit access time at noon period (min)	16.08	17.01	1.20	89.90
TAT_PMPeak	Transit access time at afternoon peak period (min)	15.81	16.93	1.10	89.90
TAT_Late	Transit access time at night period (min)	17.15	16.67	1.30	89.90

EMPIRICAL ANALYSIS

Heatmaps of taxi and Uber pick-ups

Heatmaps of taxi and Uber pick-ups are shown in *Figure 3*, to visualize the distribution of demand over space. These maps are made up of tiny dots each of them representing a single pick-up location. The bright color is caused by concentrated dots and indicates higher demand activity. *Figure 3* (a) represents activities by yellow taxis, most of which are heavily concentrated in Manhattan, as well as airports in other boroughs. Although inhabitants in Manhattan account for less than 20 percent of the total population *(2)*, Manhattan still presents significantly higher demand than other boroughs. Compared with taxis, the demand of Uber tends to be distributed more evenly throughout the city, as shown in *Figure 3* (b).



(a) Taxi (b) Uber

Figure 3 Heatmaps of taxi and Uber pick-ups in New York City.

Correlation of Taxi and Uber pick-ups

Figure 1 (a) and (b) present the taxi and Uber pick-ups by NTAs, and demands of both taxis and Uber are spatially clustered. *Figure 1* represents the Pearson coefficients of correlation between Uber and taxis using daily trip data. A high correlation between taxi and Uber pick-ups can be observed, especially in the city central areas.

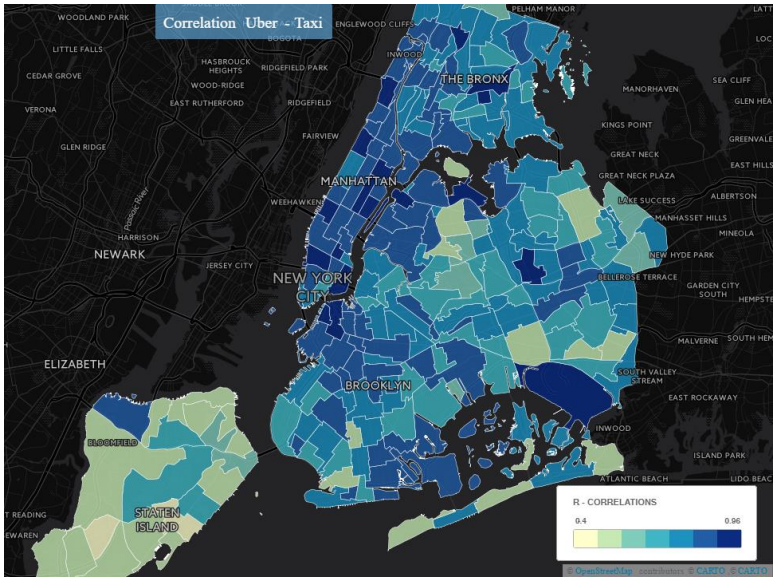


Figure 4 Pearson correlation coefficients of taxi and Uber pick-ups.

Daily Trips of Taxis and Uber

Figure 5 shows the daily trips of Uber vehicles as well as green and yellow taxis. Note that Uber data is only available from April 2014–September 2014, then from January 2015–June 2015, which explains the gap in the graph. The demand for green and yellow taxis fluctuate over time, but no significantly upward or downward trend is observed. Possibly, if we had a much longer time series, we would have been able to see some long-term patterns. The demand for green taxis shows slow but moderately steady growth. The demand for Uber shows a strongly increasing trend. Uber demand started lower than that of the green taxies, but overtook it in 2015. The pie charts on the right side of *Figure 5* show the change in market shares from 2014 to 2015. There is a dramatic drop in taxi and Uber usage on January 28 due to a winter storm.

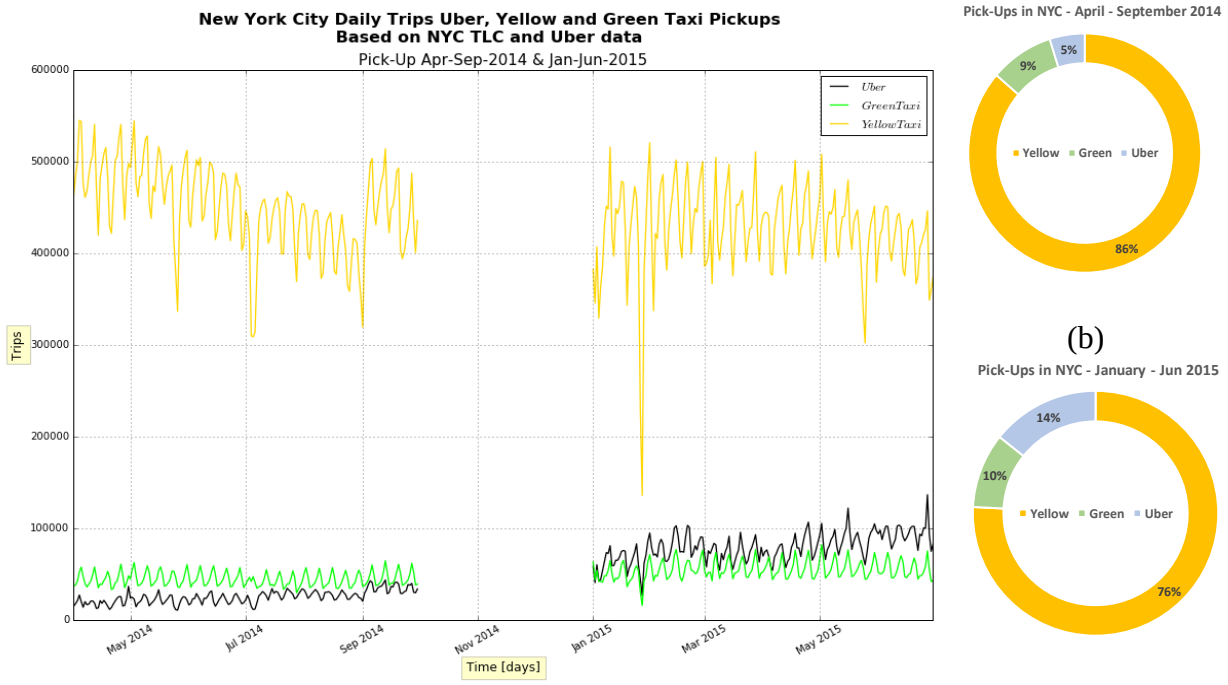
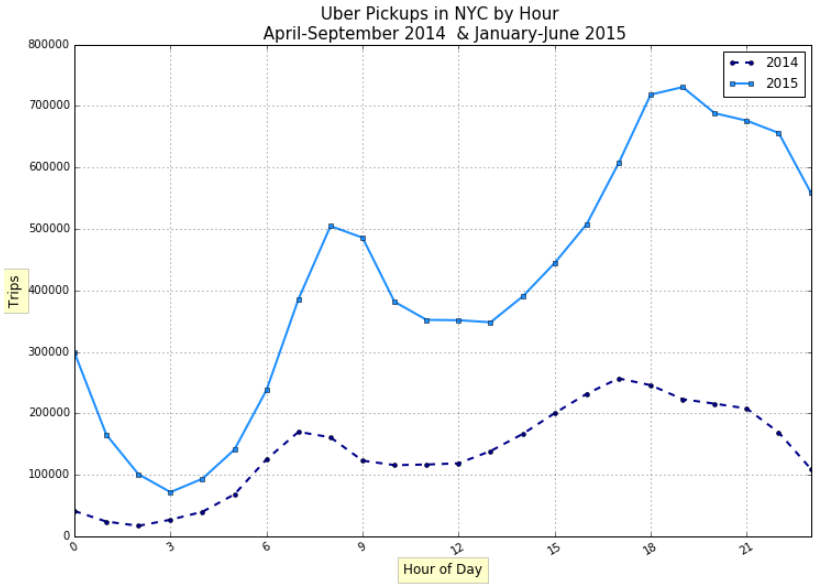


Figure 5 Daily Uber, green and yellow taxi trips (2014-2015).

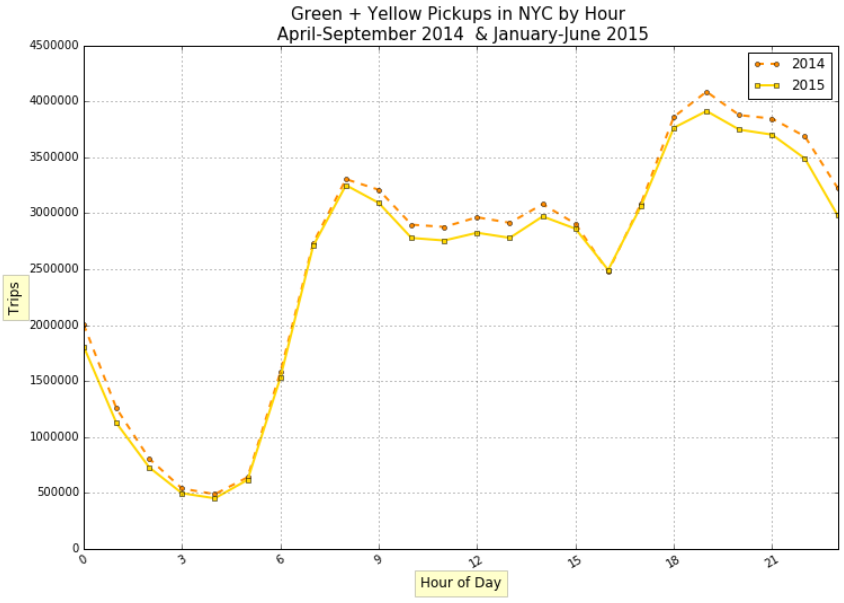
The downward trend of yellow taxi pickups coincides with the increasing trend of the Uber. The decline reflects what's happening across the country where taxi companies are regulated by local governments on everything from price to the color of their cars and location that they are allowed to operate. Meanwhile, they are trying to compete with cheaper and more flexible app-based transportation start-ups, like Uber. Other aspects that may influence the demand for traditional taxis could be the convenience factor of Uber's smartphone hailing systems, better customer service, or simply because the new ride-hailing services are cheaper mainly because in many cases the ridesharing companies are not subject to the same regulations as traditional taxis.

Distribution of Taxi and Uber Demands Over Time of Day

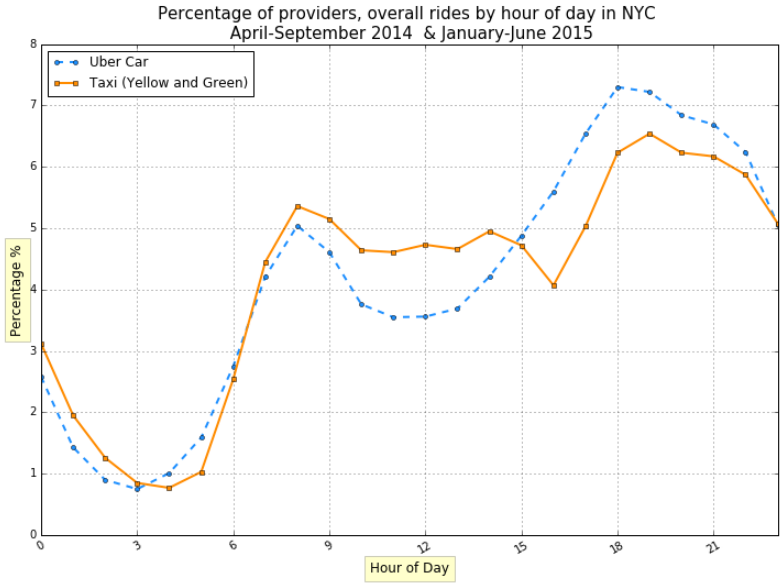
Figure 6 (a) shows a significant growth of Uber demand from 2014 to 2015, while the distribution of taxi demand over time of day remains almost the same as shown in *Figure 6* (b). In *Figure 6* (c) we plot the percentages of pick-ups per hour for both Uber and taxi. AM and PM peaks are observed in both modes, with relatively higher demands during PM peaks. The hour with the maximum demand is 19:00 for taxis and 18:00 for Uber vehicles. We can see a drop in taxi demand at 16:00 possibly because of late-afternoon shift changes of taxi drivers, while the demand for Uber keeps increasing from 12:00 till 18:00. The duration of PM peak for Uber demand is longer than that of taxi demand. The percentage of mid-day Uber demand is lower than that of taxi demand.



(a)



(b)



(c)

Figure 6 Comparison between Uber and taxi pick-ups (2014-2015).

Changes in Demand for Taxi and Uber by Boroughs

In NYC, the number of taxi/Uber pick-ups were 9.0 million more in 2015 than it did in 2014, as shown in data in *Table 2*. Demand for taxi/Uber pick-ups increase from 90.8 to 99.8 million. However, the trends for taxi/Uber are different thus, yellow taxis went down by 2.7 million pick-ups, that represent a decrease of 3.4%. Demand for green taxis instead increases by 1.8 million pick-ups, which represents an increment of 23.0%. Overall, taxi trips decrease by 0.8 million, which is 1.0% of the total trips. However, demand for Uber increased by 9.6 million additional pick-ups in 2015. Its represent an increase of 223.3 %. during this time period. Uber grew quickly during the months for which we have data.

In Brooklyn, Uber grew by 291.2%. This accounts for more than once as many pick-ups as yellow taxis, and is approaching the popularity of green taxis. In Bronx, Uber grows by 597.0% and accounts for more than three times as many pick-ups as yellow taxis, but green taxis are still more broadly used.

Manhattan has by far the largest number of taxi pick-ups. Around 92.0% of all NYC taxi pick-ups occur in Manhattan, in any given month and most of those are made by yellow taxis. Although green taxis are allowed to operate in upper Manhattan, they are taking a small fraction of yellow taxi pick-ups. Uber grew quickly during the six months for which we have data. Its drivers provided 81.0% more rides in September than in April. Uber has grown dramatically in Manhattan as well, reaching a 201.2% increase in pick-ups from 2014 to 2015, while taxi pick-ups declined by 3.0% over the same period. In Manhattan between 2014 and 2015, the number of Uber pick-ups were 6.9 million more pick-ups in 2015 than 2014, while taxis made 2.2 million fewer pick-ups during the same time period. Nevertheless, in Manhattan Uber picked up nearly 10.4 million in 2015, Uber still accounts for less than 15.0% of total Manhattan pick-ups.

Queens still has more yellow taxi pick-ups than green taxi pick-ups. That's likely because LaGuardia and JFK airports are both in Queens, and they are heavily served by yellow taxis. In Staten Island, the number of taxi pick-ups is much smaller than other boroughs and Uber represents an increase of 573.0%.

Overall, while Uber had a greater percentage of pick-ups outside of Manhattan than taxis did, there was a plenty of variation in Uber's share among neighborhoods in the outer boroughs. In NYC, there is an increment of 9 million new trips created since 2014, so that does not mean Uber is taking millions of rides away from taxis. Instead, that may be due to the fact that people are shifting from other transportation modes to ridesharing services like Uber.

Table 2 Uber, Yellow and Green Taxi Pick-Ups by Borough 2014 – 2015

Borough	Year	Yellow Taxi	Green Taxi	Uber Car
Brooklyn (BK)	2014	1895524	2681757	593597
	2015	1569702	3660406	2322000
	% Shift	-17.2%	36.5%	291.2%
Bronx (BX)	2014	54412	648259	31584
	2015	64979	707600	220146
	% Shift	19.4%	9.2%	597.0%
Manhattan (MN)	2014	72068149	2411934	3443562
	2015	69884557	2797122	10371060
	% Shift	-3.0%	16.0%	201.2%
Queens (QN)	2014	4363406	2284146	342225
	2015	4177639	2706461	1343945
	% Shift	-4.3%	18.5%	292.7%
Staten Island (SI)	2014	932	1105	1034
	2015	1087	1581	6959
	% Shift	16.6%	43.1%	573.0%
New York City (NYC)	Total 2014	78382423	8027201	4412002
	Total 2015	75697964	9873170	14264110
	Shift	-2684459	1845969	9852108
	% Shift	-3.4%	23.0%	223.3%
Total Yellow + Green + Uber Trips 2014				90821626
Total Yellow + Green + Uber Trips 2015				99835244
New trips in 2015				9013618
% Shift 2014-2015				9.9%

SPATIAL MODELING OF UBER AND TAXI DEMANDS

Spatial Dependence Test

Spatial clustering of Uber and taxi pick-ups can be observed from *Figure 1*. To quantify the spatial dependence of Uber and taxi pick-ups, the Moran's I test proposed by Moran (1948) *(23)* was conducted. Please refer to Xie et al. *(22)* for definitions of Moran's I statistics.

GeoDa (31) was used to conduct the Moran's I test. The minimum threshold distance which could ensure all the NTAs have at least one neighbor was selected *(32)*. A total of 9,999 permutations were performed to compute the pseudo p-values which indicate the significance of spatial dependence.

The estimates of Moran's I statistics for Uber and taxi pick-ups are presented in *Table 3*. It is found that both the pseudo p-values are less than 0.05, which means the spatial dependence is statistically significant at the confidence interval of 95%. If the spatial dependence is neglected, biased estimates and unreliable inferences would be obtained.

Table 3 Results of Moran's I Tests

	I	$E[I]$	$SD[I]$	z_I	Pseudo p-value
Uber	0.3684	-0.0052	0.0291	12.8213	0.0001
Taxi	0.4248	-0.0052	0.0286	15.0532	0.0001

Spatial Models

In this section, the specifications of the linear, spatial error and spatial lag models are presented. Please refer to Xie et al. *(22)* for more details on model estimation and model performance measures including maximum likelihood, AIC, and BIC.

Linear Model

A linear model assumes a linear relationship between the logarithm of demand and the vector of explanatory variables X. In matrix form, it can be expressed as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 \boldsymbol{I}) \tag{2}$$

where $\boldsymbol{y}$ is the vector of the logarithm of demand, $\boldsymbol{X}$ the vector of explanatory variables, $\boldsymbol{\beta}$ the vector of regression coefficients to be estimated and $\boldsymbol{I}$ represents the identity matrix. The error term ε is assumed to be independent and identically distributed with mean zero and constant variance.

Spatial Error Model

In the spatial error model, spatial dependence is captured via spatial error correlation. The spatial error model in matrix form can be specified as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \lambda \boldsymbol{W u} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 \boldsymbol{I}) \tag{3}$$

In *Eq. (3)*, the overall error is represented by two components namely ε, a spatially uncorrelated error term which satisfies the same assumption of error term in *Eq. (2)*, and $\boldsymbol{u}$, a spatially dependent error term. The spatial autoregressive parameter λ indicates the extent to which $\boldsymbol{u}$ of observations are correlated.

Spatial Lag Model

In the spatial lag model, spatial dependence is captured through both spatial error correlation effects and spatial spillover effects. The spatial lag model in matrix form can be specified as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \rho\boldsymbol{W}\boldsymbol{y} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2\boldsymbol{I}) \tag{4}$$

where $\rho\boldsymbol{W}\boldsymbol{y}$ is a spatially lagged dependent variable, ρ is a spatial autoregressive parameter, and the rest notation is as before. The assumption of error term ε is the same as the one in *Eq. (2)*. Similar to λ in *Eq. (3)*, ρ denotes the spatial correlation of observations.

Modeling Results

The linear model, spatial error model and spatial lag model specified in the last section were developed for the taxi and Uber demands. Stepwise AIC was used to select variables. All the explanatory variables included in the three models were kept the same to conduct effective comparisons of different model structures. The coefficient estimations and statistic measures of these three models were reported in *Table 4*. A broadly used statistic indicator p-value was used to test the significance of explanatory variables. Most of the explanatory variables were found to be statistically significant at 95% level (p-values<0.05).

The autoregressive parameters λ in the spatial error model and ρ in the spatial lag model are found to be highly significant, and these are reported in *Table 4*. This result along with the Moran's I statistics afford strong evidence and suggest the spatial dependence between taxi/Uber pick-ups. However, the use of the parameter R^2 should be used with attention in spatial models whose residuals are not independent.

Important criterions are also presented to completely evaluate the three spatial models as Likelihood-based criteria such as $LL_{\max}$, AIC and BIC. Compared with the linear models, both the spatial lag and spatial error models show significant improvement (AIC and BIC differences are greater than 10). There is a significant difference between linear and spatial models by accommodating spatial autoregressive processes. In the taxi model, the spatial lag model is better than the spatial error model (AIC difference is more than 4, that mean the two models can be regarded as a considerable difference). However, the BIC parameter difference is 0.27 (less than 2), not worth more than a bare mention. In the Uber model, the result is less evident. However, we can say that the spatial lag model is slightly better than the spatial error model, since AIC difference is 1.5 (less than 2). This finding is consistent with the results of the $LL_{\max}$ criteria as well. The spatial lag models outperform the others by considering both spatial spillover and spatial error correlation effects.

1 **Table 4 Model Results**

(a) Model Assessment

Model Assessment	Taxi Model			Uber Model		
	Linear	Spatial Error	Spatial Lag	Linear	Spatial Error	Spatial Lag
R-Squared	0.80	0.87	0.87	0.79	0.83	0.82
Likelihood	-339.33	-306.26	-303.49	-279.48	-265.32	-263.56
AIC	700.65	634.53	630.98	582.96	554.63	553.12
BIC	736.65	670.53	670.26	622.23	593.91	595.67

(b) Coefficient Estimates (Taxi)

Variable	Linear		Spatial Lag		Spatial Error	
	Estimate	p-Value	Estimate	p-Value	Estimate	p-Value
Constant	5.5308	<0.0001	1.5166	0.0056	5.5336	0.0017
Brooklyn	3.8031	<0.0001	1.2903	0.0035	4.0933	0.0244
Manhattan	5.1014	<0.0001	1.7398	0.0040	4.2160	0.0215
Queens	4.4335	<0.0001	1.7966	0.0000	3.5109	0.0540
Bronx	3.9621	<0.0001	1.4147	0.0014	3.1861	0.0873
TAT	-0.0349	<0.0001	-0.0183	0.0063	-0.0326	0.0010
BusStop	0.0049	0.16859	0.0066	0.0189	0.0070	0.0232
RoadsMi	0.0177	0.0003	0.0159	0.0000	0.0223	<0.0001
VehOwn	-0.4556	<0.0001	-0.2721	<0.0001	-0.2420	<0.0001
TotJob	0.1085	<0.0001	0.0722	0.0000	0.0457	0.0118
EduBac	0.0959	<0.0001	0.0492	0.0024	0.0561	0.0014
Autoregressive Parameter						
ρ			0.6115	<0.0001		
λ					0.8669	<0.0001

(c) Coefficient Estimates (Uber)

Variable	Linear		Spatial Lag		Spatial Error	
	Estimate	p-Value	Estimate	p-Value	Estimate	p-Value
Constant	4.0076	<0.0001	1.4741	0.0100	4.1250	<0.0001
Brooklyn	4.0021	<0.0001	2.1785	<0.0001	4.2380	<0.0001
Manhattan	4.0685	<0.0001	2.2302	<0.0001	4.1738	<0.0001
Queens	3.9632	<0.0001	2.3671	<0.0001	4.3705	<0.0001
Bronx	2.9271	<0.0001	1.6958	<0.0001	2.8408	0.0006
TAT	-0.0206	0.0009	-0.0162	0.0031	-0.0302	0.0001
BusStop	0.0078	0.0039	0.0073	0.0017	0.0063	0.0142
RoadsMi	0.0087	0.0154	0.0088	0.0055	0.0148	0.0001
VehOwn	-0.2375	<0.0001	-0.1521	<0.0001	-0.1378	0.0006
TotJob	0.0653	0.0004	0.0476	0.0028	0.0279	0.0940
CapInc	0.0146	0.0001	0.0103	0.0019	0.0096	0.0043
EduBac	0.0000	0.1069	0.0179	0.2983	0.0323	0.0592
Autoregressive Parameter						
ρ			0.4607	<0.0001		
λ					0.7124	<0.0001

The spatial lag models are used to assess the effects of variables presented in the following discussion. A study by Xie et al. *(33)* showed the practical use in interpreting variables. In interpreting the signs of the coefficients in *Table 4,* a positive sign implies an expected increase in demand, while a negative sign suggests an expected decrease. The exponents of coefficients provide an intuitive and quantitative way to indicate the percentage change in the dependent variable with respect to a unit change of the explanatory variables *(34).*

Results show the statistically significant variables namely, *Brooklyn*, *Manhattan*, *Queens*, *Bronx*, *EducBac*, *TotJob*, *RoadsMi*, and *BusStop*, are positively related to taxi and Uber demand. On the other hand, *TAT* and *VehOwn* are negative related. The coefficient of transit access time (*TAT*) is negative, meaning that neighborhoods with better transit accessibility also have more taxi/Uber pick-ups. One-unit decrease in *TAT* is associated with 1.8% ($e^{0.0183}$-1) and 1.6% ($e^{0.0162}$-1) increase in taxi and Uber pick-ups, respectively. In addition, one-unit increase in *BusStop* is associated with 0.7% ($e^{0.0066}$-1) and 0.7% ($e^{0.0073}$-1) increase in taxi and Uber pick-ups, respectively. This finding seems to be counter-intuitive. A possible reason is that neighborhoods with lower *TAT* are mostly in city central areas like Manhattan, which have higher travel demand. So actually, lower *TAT* is not the cause for higher Uber/taxi trips. They are both correlated with the high travel demand.

The demand for taxi and Uber are positively related to the length of roadways (*RoadsMi*). This is because Taxi and Uber's demands are also related to the roadway density. Demand have a significant impact in areas with longer mileage of roadway. The coefficient of vehicle ownership (*VehOwn*) is negative, meaning that people who own a car are less likely to use taxi or Uber. Zones with a higher rate of car ownership, use less taxi and Uber services, which is consistent with the expected relationships. One unit decrease in *VehOwn* is associated with -23.8% ($e^{-0.2721}$-1) and -14.1% ($e^{-0.1521}$-1) increase in taxi and Uber pick-ups, respectively.

Socioeconomic factors like income (*CapInc*), higher education (*EduBac*), and the total number of jobs (*TotJob*) are positively related to the number of taxi and Uber pick-ups as expected. Normally, the areas of higher average income in NYC are located either uptown Manhattan, or some areas of Brooklyn, where residents prefer to use taxis for travel. Job and education opportunities are also related to the level of economic activities, and more activities require more taxi and Uber service. One unit increase in *EduBac* is associated with 5.0% ($e^{0.0024}$-1) and 1.8% ($e^{0.0179}$-1) increase in taxi and Uber pick-ups, respectively. Similarly, one unit increase in *TotJob* is associated with 7.5% ($e^{0.0159}$-1) and 4.9% ($e^{0.0088}$-1) increase in taxi and Uber pick-ups, respectively.

CONCLUSIONS

This study develops Neighborhood Tabulated Area (NTA) based demand forecasting models, using large-scale Uber and taxi pick-ups data from New York City. Major contributions of this paper compared with previous studies are the use of Uber data and the consideration of spatial dependence in modeling taxi and Uber's demand. In the empirical analysis, this study demonstrated the temporal and spatial variation of taxi and Uber's demands. The relationship between taxi demand and transit accessibility and other socio-economic and transportation-related factors is also fully explored using spatial models.

In the empirical analysis, we explore the spatio-temporal patterns of Uber and taxi pick-up data. A high correlation between taxi and Uber pick-ups can be observed, especially in the city central areas. Compared with taxis, the demand of Uber tends to be distributed more evenly

throughout the city. The duration of PM peak for taxi demand is shorter than that of Uber demand, due to the late-afternoon shift changes of taxi drivers. From 2014 to 2015, Uber trips increased dramatically by 10 million (223.3%), while taxi trips (include both yellow and green taxis) decreased slightly by 0.8 million (1.0%). The rate of growth of Uber is the lowest in Manhattan (201.2%), and the highest in the outer boroughs like Bronx (597.0 %) and Staten Island (573.0%).

Spatial dependence of pick-up data was investigated specifically in this study. Moran's I test was conducted to explore the spatial dependence of taxi and Uber pick-ups. It is confirmed that taxi and Uber demands were significantly correlated spatially at the confidence interval of 95%. Linear models, spatial error models, and spatial lag models are developed to estimate the taxi and Uber demands of each neighborhood using socio-economical and transportation-related characteristics. The spatial lag model is found to perform better than the others by capturing the spatial dependence via a spatially lagged dependent variable. Key variables affecting taxi and Uber demands include transit access time (*TAT*), length of roadways, vehicle ownership, education, employment, and income. Neighborhoods with lower transit access time (*TAT*), higher length of roadways, lower vehicle ownership, higher income and more job opportunities are associated with higher taxi/Uber demands. One-unit decrease in *TAT* is associated with 1.8% and 1.6% increase in taxi and Uber pick-ups, respectively. This finding seems to be counter-intuitive. A possible reason is that neighborhoods with lower *TAT* are mostly in city central areas like Manhattan, which have higher travel demand.

For future study, the spatial lag model will be integrated with other techniques such as Bayesian networks. Improved versions of this spatial model will provide alternative ways of calculating *TAT* using the complete GTFS bus schedule since only the subway data was considered in this study. It is acceptable to calculate *TAT* using subway data in Manhattan considering high density of subway stations, but this approach may have bias in the outer boroughs.

Since Uber and taxi services are competing on travel demands in the real world, the inter-relationship between Uber and taxi demands can be considered in a unified model. It is also possible to model the interaction between taxi demand and the demand for all other alternative transportation modes such as bike sharing and other sharing services like Lyft.

We also need to model the interaction between taxi usage and the use of all other alternative transportation modes such as, like bike sharing and other sharing service like Lyft that might also have an impact on taxi usage. However, obtaining micro-level data for Lyft and other ride-sharing services remains to be a major challenge for studies such as this one. These new companies are highly encouraged to provide more data to researchers to enable transportation community to plan better for the future. Considerations like if Uber is being used for first-mile/last-mile problems as well as are pick-ups/drop-offs clustered near subway stations in the outer boroughs will be the focus of future research.

1 **REFERENCES**

1. Uber, *Chicago case study*, 2015. https://uber-static.s3.amazonaws.com/web-fresh/legal/Uber_Chicago_CaseStudy.pdf. Accessed March 20, 2016.
2. U.S. Census Bureau, *U.S Census 2010 Interactive Population Map*. http://www.census.gov/2010census/popmap. Accessed April 20, 2016.
3. Schaller Consulting, *The New York City Taxicab Fact Book*, 2006. Available on-line at http://www.schallerconsult.com/taxi/taxifb.pdf. Accessed on February 20, 2016.
4. De Blasio, B., Joshi, M., 2016. *2016 Taxicab Fact Book. City of New York*. http://www.nyc.gov/html/tlc/downloads/pdf/2016_tlc_factbook.pdf. Accessed April 5, 2016.
5. Joshi, M., 2016. *Your guide to Boro Taxi City of New York*. http://www.nyc.gov/html/tlc/html/passenger/ shl_passenger.shtml. Accessed May 3, 2016.
6. Joshi, M., 2015. *Hail Market Analysis City of New York.* http://www.nyc.gov/html/tlc/downloads/pdf/hail_market_analysis_2015.pdf. Accessed April 20, 2016.
7. King, D. A., Peters, J. R., and Daus, M. W. Taxicabs for Improved Urban Mobility: Are We Missing an Opportunity? Transportation Research Board - 91st Annual Meeting, 37 Washington, D.C. Transportation Research Board of the National Academies, 2012.
8. Qian, X. and Ukkusuri, S. V. Spatial variation of the urban taxi ridership using GPS data. Applied Geography, Vol. 59, 2015, pp. 31-42.
9. Chang, H.-w., Tai, Y.-c., and Hsu, J. Y.-j. Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, Vol. 5, 1, 2010, pp. 3-18.
10. Moreira-Matias, L., Gama, J., Ferreira, M., and Damas, L. A Predictive Model for the Passenger Demand on a Taxi Network. *15th International IEEE Conference on Intelligent. Transportation Systems Anchorage,* Alaska, USA. IEEE, 2012.
11. Phithakkitnukoon, S., Veloso, M., Bento, C., Biderman, A., and Ratti, C. Taxi-Aware Map: Identifying and Predicting Vacant Taxis in the City. In B. d. Ruyter, R. Wichert, D. V. Keyson et al. (Eds.), Ambient Intelligence, LNCS 6439, Springer, Berlin, 2010, pp. 86-95.
12. Morgul, E. F., Ozbay, K., Iyer, S., and Holguín-Veras, J. Commercial Vehicle Travel Time Estimation in Urban Networks using GPS Data from Multiple Sources. Transportation Research Board - 92nd Annual Meeting, Washington, D.C. Transportation Research Board of the National Academies, 2013.
13. Austin, D., and Zegras, C. Taxicabs as public transportation in Boston, Massachusetts. Transportation Research Record, No. 2277, 2012, pp. 65-74.
14. Yang, C., Gonzales, E. Modeling Taxi Demand and Supply in New York City Using Large-Scale Taxi GPS Data. Seeing Cities Through Big Data - Research, Methods and Applications in Urban Informatics (In Press).
15. Yang, C., Morgul, E. F., Gonzales, E. J., and Ozbay, K. Comparison of Mode Cost by Time of Day for Nondriving Airport Trips to and from New York City's Pennsylvania Station. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2449, 2014, pp. 34-44.
16. Parfenov, S., Weeks, A., and Alam, Z. Travel Patterns of NYC's Yellow Taxis: *Routing*,

Activity and Results ESRI International User Conference, San Diego, CA, 2014.
17. Ferreira, N., Poco, J., Vo, H. T., Freire, J., and Silva, C. T. Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. IEEE Transactions on Visualization and Computer Graphics, Vol. 19, 12, 2013, pp. 2149-2158.
18. Yang, C. and Gonzales, E. J. Modeling taxi trip demand by yime of day in New York City. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2429, 2014, pp. 110-120.
19. Yang, C. and Gonzales, E. J., 2016. Modeling the spatial variation of taxi trip demand and Supply in New York City.
20. Anselin, L., 1988b. Spatial Econometrics: Methods and Models. Springer.
21. Xie, K., Wang, X., Ozbay, K., Yang, H., 2014b. Crash frequency modeling for signalized intersections in a high-density urban road network. Anal. Methods Accid. Res. 2, 39–51.
22. Xie, K., Ozbay, K., Yang, H., 2014b. Spatial analysis of highway incident durations in the context of Hurricane Sandy. Accident Analysis and Prevention 74 (2015) 77–86.
23. Moran, P.A., 1948. *The interpretation of statistical maps. Journal of the Royal Statistical Society*. Series B (Methodological) 10(2), pp. 243-251.
24. NYC Department of city Planning. *Neighborhood Tabulated Areas NTA.* http://www1.nyc.gov/site/planning/data-maps/open-data.page. Accessed March 18, 2016.
25. U.S. Census Bureau, *American Fact Finder*. http://factfinder.census.gov
26. *NYC Taxi and Limousine Commission (TLC).* http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml. Accessed March 8, 2016.
27. Schneider T.W. *Git-Hub Repo*. https://github.com/toddwschneider/nyc-taxi-data. Accessed March 8, 2016.
28. Donnelly, F. Introduction to the NYC Geodatabase (nyc_gdb) ArcGIS Version. http://www.baruch.cuny.edu/geoportal/nyc_gdb/data/intro_nycgdb_arc.pdf. Accessed May 30, 2016
29. The Metropolitan Transportation Authority MTA. http://web.mta.info/developers/developer-data-terms.html#data. Acceded on May 28 of 2016.
30. Browning, R., E. Baker, J. Herron, and R. Kram. Effects of obesity and sex on the energetic cost and preferred speed of walking. *Journal of Applied Physiology*, Vol. 100, 2005, pp. 390-398.
31. Anselin, L., 2003. GeoDa™ 0.9 user's guide. Urbana 51, pp. 61801.
32. Mitchell, A., 2005. The ESRI Guide to GIS Analysis, Volume 2: Spatial Measurements and Statistics. ESRI Press.
33. Xie, K., X. Wang, H. Huang, and X. Chen, 2013. Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accident Analysis and Prevention 50, pp. 25-33.
34. Tavassoli Hojati, A., Ferreira, L., Washington, S., Charles, P., 2013. Hazard based models for freeway traffic incident duration. Accid. Anal. Prev. 52, 171–181.